Course: MMDS

Lab 02

HDFS

Fill answers of the questions below in the given tables.
Your screenshots must contain commands for required operations.

Question 1:

Download and unzip lab02.zip to obtain 365 text files.
Create the folder **/user/<username>/lab02** in your HDFS
Copy the entire **lab02/** from your local filesystem to **/user/<username>/lab02** in HDFS
Take two screenshots to show your results.
For example,

```
Found 1 items
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:27 /user/ntan/lab02/lab02
```

```
Found 365 items
-rw-r--r--   1 ntan supergroup          1 2021-05-18 15:27 /user/ntan/lab02/lab02/2021_01_01.txt
-rw-r--r--   1 ntan supergroup          2 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_02.txt
-rw-r--r--   1 ntan supergroup          2 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_03.txt
-rw-r--r--   1 ntan supergroup          1 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_04.txt
-rw-r--r--   1 ntan supergroup          2 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_05.txt
-rw-r--r--   1 ntan supergroup          2 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_06.txt
-rw-r--r--   1 ntan supergroup          1 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_07.txt
```

My screenshot

```
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -ls /user/pkucpkam/lab02
Found 1 items
drwxr-xr-x   - pkucpkam supergroup          0 2025-01-18 12:42 /user/pkucpkam/lab02/lab02
```

```
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -ls /user/pkucpkam/lab02/lab02
Found 365 items
-rw-r--r--   2 pkucpkam supergroup          1 2025-01-18 12:41 /user/pkucpkam/lab02/lab02/2021_01_01.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_02.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/lab02/2021_01_03.txt
-rw-r--r--   2 pkucpkam supergroup          1 2025-01-18 12:41 /user/pkucpkam/lab02/lab02/2021_01_04.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_05.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_06.txt
-rw-r--r--   2 pkucpkam supergroup          1 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_07.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/lab02/2021_01_08.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/lab02/2021_01_09.txt
-rw-r--r--   2 pkucpkam supergroup          1 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_10.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_11.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/lab02/2021_01_12.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_13.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/lab02/2021_01_14.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/lab02/2021_01_15.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_16.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/lab02/2021_01_17.txt
```

Question 2:

Create 12 folders, corresponding to 12 months in a year, in **/user/<username>/lab02** in HDFS.
Notice: folder names must in the format like _01, _02, …, _12
Take a screenshot to show your result.
Hint:

- loop in bash scripts
- $(printf …) for formatting strings in bash scripts.

For example,

```
Found 13 items
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_01
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_02
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_03
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_04
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_05
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_06
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_07
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_08
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_09
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_10
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_11
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:30 lab02/_12
drwxr-xr-x   - ntan supergroup          0 2021-05-18 15:27 lab02/lab02
```

My screenshot

```
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ nano create_folders.sh
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ ./create_folders.sh
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ hadoop fs -ls /user/pkucpkam/lab02
Found 13 items
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_01
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_02
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_03
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_04
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_05
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_06
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_07
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_08
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_09
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_10
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_11
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:10 /user/pkucpkam/lab02/_12
drwxr-xr-x   - pkucpkam supergroup          0 2025-02-11 12:03 /user/pkucpkam/lab02/lab02
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$
```

Question 3:

Text files in **lab02/** have a filename format like **YYYY_MM_DD.txt**
Move each text file in **lab02/** to the folder which is corresponding to the month in its filename. E.g. **20201_01_01.txt** to **_01/**
Take a screenshot to show the content in **lab02/_01/**
Take a screenshot to show the sizes (in bytes) of the 12 folders.
Hint: loop in bash scripts
For example,

```
Found 31 items
-rw-r--r--   1 ntan supergroup          1 2021-05-18 15:27 lab02/_01/2021_01_01.txt
-rw-r--r--   1 ntan supergroup          2 2021-05-18 15:26 lab02/_01/2021_01_02.txt
-rw-r--r--   1 ntan supergroup          2 2021-05-18 15:26 lab02/_01/2021_01_03.txt
-rw-r--r--   1 ntan supergroup          1 2021-05-18 15:26 lab02/_01/2021_01_04.txt
-rw-r--r--   1 ntan supergroup          2 2021-05-18 15:26 lab02/_01/2021_01_05.txt
```

```
58  58  lab02/_01
52  52  lab02/_02
58  58  lab02/_03
56  56  lab02/_04
58  58  lab02/_05
56  56  lab02/_06
58  58  lab02/_07
58  58  lab02/_08
56  56  lab02/_09
58  58  lab02/_10
56  56  lab02/_11
58  58  lab02/_12
```

My screenshot

```
pkucpkam@PhucPham:~$ hadoop fs -ls /user/pkucpkam/lab02/_01
Found 31 items
-rw-r--r--   2 pkucpkam supergroup          1 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_01.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_02.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_03.txt
-rw-r--r--   2 pkucpkam supergroup          1 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_04.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_05.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_06.txt
-rw-r--r--   2 pkucpkam supergroup          1 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_07.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_08.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_09.txt
-rw-r--r--   2 pkucpkam supergroup          1 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_10.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_11.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_12.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_13.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_14.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_15.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_16.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_17.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:42 /user/pkucpkam/lab02/_01/2021_01_18.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_19.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_20.txt
-rw-r--r--   2 pkucpkam supergroup          2 2025-01-18 12:41 /user/pkucpkam/lab02/_01/2021_01_21.txt
```

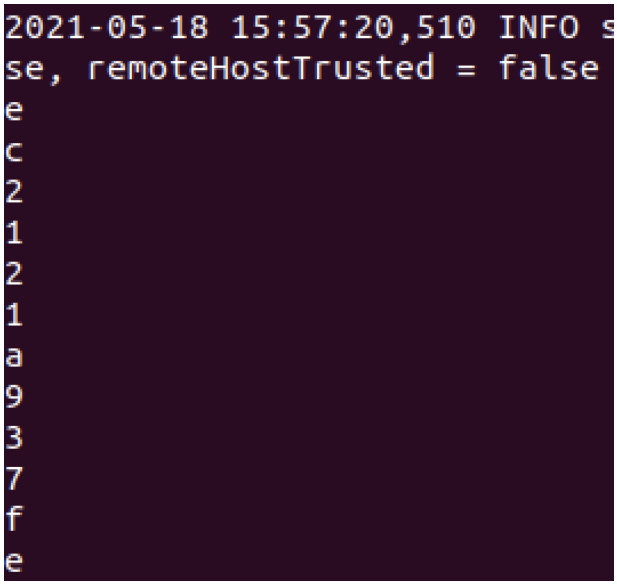
```
58 58 /user/pkucpkam/lab02/_01
52 52 /user/pkucpkam/lab02/_02
58 58 /user/pkucpkam/lab02/_03
56 56 /user/pkucpkam/lab02/_04
58 58 /user/pkucpkam/lab02/_05
56 56 /user/pkucpkam/lab02/_06
58 58 /user/pkucpkam/lab02/_07
58 58 /user/pkucpkam/lab02/_08
56 56 /user/pkucpkam/lab02/_09
58 58 /user/pkucpkam/lab02/_10
56 56 /user/pkucpkam/lab02/_11
58 58 /user/pkucpkam/lab02/_12
```

Question 4:

Using **cat** to display the content of all files (in month order)

- day 10th
- day 01st
- day 07th
- day 04th

Take 4 screenshots of the corresponding results above.
For example,



My screenshot
day 10th

```
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ hdfs dfs -cat /user/pkucpkam/lab02/_*/????_??_10.txt | sort -t'/' -k3,3 -k1,1 | tr -d '\n' | awk '{s
plit($0, chars, ""); for (i=1; i<=length($0); i++) print substr($0, i, 1)}'
2025-02-11 12:32:48,979 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
_
<
3
_
_
i
_
_
l
o
v
e
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ S
```

day 1st

```
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ hdfs dfs -cat /user/pkucpkam/lab02/_*/????_??_01.txt | sort -t'/' -k3,3 -k1,1 | tr -d '\n' | awk '{s
plit($0, chars, ""); for (i=1; i<=length($0); i++) print substr($0, i, 1)}'
2025-02-11 12:34:03,345 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
_
_
y
o
u
_
_
f
o
r
_
_
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$
```

day 07th

```
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ hdfs dfs -cat /user/pkucpkam/lab02/_*/????_??_07.txt | sort -t'/' -k3,3 -k1,1 | tr -d '\n' | awk '{s
plit($0, chars, ""); for (i=1; i<=length($0); i++) print substr($0, i, 1)}'
2025-02-11 12:34:43,354 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
a
_
_
t
h
o
u
s
a
n
d
_
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$
```

day 04th

```
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$ hdfs dfs -cat /user/pkucpkam/lab02/_*/????_??_04.txt | sort -t'/' -k3,3 -k1,1 | tr -d '\n' | awk '{s
plit($0, chars, ""); for (i=1; i<=length($0); i++) print substr($0, i, 1)}'
2025-02-11 12:35:03,359 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
_
y
e
a
r
s
_
_
<
3
_
.
pkucpkam@PhucPham:~/Desktop/hadoop-3.2.1$
```

Question 5:

Concatenate contents in the 4 screenshots of Question 4 to form the completed message. Finally, write it in the given table.

My answer

_<3__i__love__you__for__a__thousand__years__<3_.